A

Topic Approval Seminar Report

on

Heatwave Monitoring and Prediction for Sustainable Healthcare using Incremental Machine Learning

Submitted in partial fulfilment of the requirement for the degree of Ph.D.

in

Science and Technology

(Information Technology)

By

Sulochana Devi
Admission/ Registration No: 2220220398

Under the guidance of:

Dr. Radhika Kotecha
Professor and Head - Department of Information Technology

Department of Information Technology
K J Somaiya Institute of Technology
Somaiya Ayurvihar Complex Eastern Express Highway Near Everard Nagar, Sion East, Mumbai, Maharashtra 400022

मुंबई विद्यापीठ
University of Mumbai

2023-24

CERTIFICATE OF APPROVAL

This is to certify that **Sulochana Devi** has delivered the Topic Approval Seminar on **"Heatwave Monitoring and Prediction for Sustainable Healthcare using Incremental Machine Learning"** on **February 13, 2024** and submitted the report to **K. J. Somaiya Institute of Technology, Sion, Mumbai,** in the partial fulfilment of the degree of **Ph.D.** in **Science and Technology (Information Technology)**.

13/2/24

Dr. Radhika Kotecha

Guide

13/02/24

Dr. Hariram Chavan

Internal Examiner

Dr. Kamal Shah

External Examiner

13/2/24

Dr. Radhika Kotecha

Head of Department

13/02/24

Dr. Hariram Chavan

Ph.D. Coordinator

01/03/24

Dr. Vivek Sunnapwar

Principal

Principal
K J Somaiya Institute of Technology
Sion (E), Mumbai - 400 022

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evokepenal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Sulochana Devi

Registration No: 2220220398

Date:13/02/2024

Place: Mumbai

ACKNOWLEDGEMENT

I am honored to present "**Heatwave Monitoring and Prediction for Sustainable Healthcare using Incremental Machine Learning**" as my topic approval seminar. I would like to express my heartfelt thanks to my supervisor and our Head of Department **Dr. Radhika Kotecha** for her invaluable guidance, continuous support and valuable suggestions throughout the seminar report.Her expertise, knowledge and valuable insights have greatly enriched understanding of the subject matter.

I express my deepest gratitude towards our Ph. D. coordinator **Dr. Hariram Chavan** for his valuable and timely advice during the various phases in our Ph. D. course work. I am using this opportunity to express my profound gratitude to our principal **Dr. Vivek Sunnapwar** for providing us with all properfacilities, support and motivation.

I wish to thank the India Meteorological Department (IMD) for providing me with the data.

I would like to express my sincere gratitude and appreciation to all those who have contributed to the successful completion of my seminar report. Without their support and guidance, this endeavor would not have been possible.

Sulochana Devi

Registration No: 2220220398

Abstract

The rise in global temperatures, coupled with the escalating frequency and intensity of heatwaves, stands as an alarming consequence of climate change, attracted significant attention within the realms of Sustainable Development Goals 3 and 13 by the United Nations. This escalating trend poses a significant threat to human health, as prolonged exposure to heat undermines the body's natural ability to regulate its internal temperature. Consequently, this vulnerability sets off a chain reaction of health issues, ranging from mild conditions like heat cramps and heat exhaustion to more severe complications such as heatstroke, hyperthermia, and even fatalities. In recent times, the impact of these heatwaves has been sharply felt by large populations, giving rise to critical public health emergencies. Vulnerable populations, including the elderly and those with pre-existing health conditions, face increased risks during the events of heat wave, further burdening healthcare systems and resources. Early prediction of heatwaves can enable the provision of crucial alerts to the public and empower authorities to implement targeted measures, proactively mitigating health hazards associated with extreme heat events. Early prediction of heatwaves can provide crucial alerts to the public, prompting them to take self-protective measures. It also empowers authorities to implement targeted interventions, proactively mitigating health hazards associated with extreme heat events.

This research endeavors to create a sustainable early warning system, employing incremental machine learning on seven decades of historical meteorological data from the India Meteorological Department. The goal is to predict probable heatwaves over the next seven days as well as identify locations prone to heatwaves. The emphasis lies in the dynamic and continuous learning process of incremental machine learning, enhancing the accuracy of predictions. With thorough implementation and performance evaluation, the work aims to preempt heat-related health emergencies, promote proactive measures to enhance community resilience to heatwave challenges, and ultimately contribute to sustainable healthcare.

Keywords: Heatwave, Sustainable Healthcare, Incremental Machine Learning, Prediction, Meteorological data.

Table of Contents

Acknowledgement ..... **iv**

Abstract… ..... **v**

Table of Contents… ..... **vi**

List of Figures… ..... **vii**

1. Introduction **1**

1.1 Motivation… 1

1.2 Problem Statement 2

1.3 Objective 3

1.4 Scope … 3

2. Related Work **4**

2.1 AI Bases Forecasting Methods 4

2.1.1 Long-Short-Term Memory (LSTM) 4

2.1.2 Random Forest… 6

2.1.3 Support Vector Machine (SVM) 7

2.2 Literature Survey… 8

2.2.1 International Status 9

2.2.2 National Status 13

3. Proposed Methodology **15**

4. Expected Contribution & Future Plan **17**

5. References… **18**

List of Figures

Figure No.	Title	Page No.
1	LSTM Network Architecture	5
2	Random Forest	6
3	Main Concept of SVM	8
4	Proposed Approach for Heatwave Prediction	15
5	Timeline Chart	17

1. Introduction

1.1 Motivation

In recent years, there has been a concerning increase in the occurrence and severity of heatwaves, marking them as some of the most destructive forms of extreme weather [1]. Heatwaves have become some of the worst extreme weather events recently, causing big problems for people, farming, and nature. Besides their main impacts, heatwaves can make other bad events worse, especially when there's also a lack of rain. This makes wildfires more likely, kills plants, and ruins crops. Because people are changing the climate, these dangerous events are happening more often, lasting longer, and becoming stronger over the last few decades. According to a report by a group called the Intergovernmental Panel on Climate Change (IPCC), the number of heatwaves has doubled since the middle of the 20th century, and they've gotten much worse too [2].

The impact of heatwaves is widespread, affecting many countries including India. In north-western India, temperatures are getting hotter, ranging from 36-38°C on average, and there's a significant increase in extremely hot days, with temperatures above 35°C [3]. A study by the Tata Centre for Development (TCD) at the University of Chicago, USA, suggests that by 2100, around 1.5 million more people in India could die each year due to extreme heat [4].

Heatwaves are characterized by magnitude, duration, extent, severity, and timing in the heat season, with additional measures like onset date, minimum night time temperatures, heatwave days, number, and frequency. A heat index reflects how temperature feels to the human body, considering factors like humidity, sun exposure, and wind speed. Heat Index and Humidex consider temperature and humidity, while Wet-bulb Globe Temperature (WBGT) considers temperature, humidity, solar radiation, and wind speed [5].

Characterized by prolonged periods of abnormally high temperatures, heatwaves pose severe challenges to human health, agriculture, energy systems, and ecosystems. Vulnerable populations, such as the elderly, children, and those with pre-existing health conditions, are particularly susceptible, placing an immense burden on healthcare systems and resources during heatwave episodes. The impacts of heatwaves extend far beyond the immediate heat-related illnesses and fatalities; they disrupt daily life, strain healthcare facilities, and exacerbate existing vulnerabilities, particularly in urban areas [6], [7].

With climate change projections indicating further escalation of heatwave occurrences, the need for effective prediction and preparedness measures has never been more urgent. Meteorological data plays a fundamental role in understanding the complex dynamics of heatwave events. Historical weather information, encompassing parameters like temperature, humidity, and wind speed, offers valuable insights into the factors driving the formation and duration of heatwaves. Heatwaves, with their escalating frequency and intensity, have

devastating impacts, making preparedness essential at different lead times to effectively mitigate their effects. [8].

The increasing complexities and non-linear nature of heatwave dynamics have challenged the accuracy and lead time of traditional prediction methods, which often relied on historical meteorological data and statistical models. Though existing meteorological prediction systems have achieved certain accuracy levels, they have limits in adaptability and the volume of data they can process [9].

In recent years, the rapid advancement of Artificial Intelligence (AI) has shown great promise in transforming various scientific domains. AI techniques, including machine learning algorithms, neural networks, and deep learning architectures, have gained considerable traction in weather forecasting and climate modelling due to their ability to handle vast amounts of data and identify subtle patterns that traditional methods might miss. AI-driven approaches offer new avenues for improving heatwave prediction and mitigating their impacts. The adaptability and capacity of AI to identify intricate patterns within vast datasets can greatly enhance the timeliness and precision of predicting heatwaves. AI algorithms in heatwave prediction can use machine learning and deep learning models to analyse historical meteorological data and climate patterns, identifying indicators of heatwaves [10].

This research seeks to deepen our understanding of the underlying meteorological drivers of heatwaves. By developing a robust heatwave prediction model based on meteorological data, this research aims to contribute to climate resilience and preparedness efforts, providing valuable insights to aid in the mitigation of heatwave impacts on vulnerable communities.

1.2 Problem Statement

In response to the escalating challenges posed by increasing heatwaves and their adverse impact on public health, this research addresses the critical need for proactive measures in sustainable healthcare. By leveraging a comprehensive dataset spanning the past 70 years provided by the Indian Meteorological Department (IMD), the work aims at development of an Incremental Machine Learning system for Heatwave Monitoring and Prediction up to seven upcoming days. By integrating real-time meteorological data with long-term trends, the goal is to pre-empt heat-related health emergencies by forecasting heatwaves and their locations PAN India with precision, and to refine predictive models over time through incremental learning.

The proposed system aspires to serve as an early warning tool, aiding healthcare professionals and policymakers in implementing targeted interventions to mitigate the consequences of health emergencies due to high temperatures. This proactive approach aligns with the principles of sustainable healthcare, ensuring the well-being of communities by harnessing technological advancements to respond effectively to the evolving challenges posed by climate-induced heatwaves.

1.3 Objectives

The work aims to develop more accurate and reliable methods to forecast and predict the occurrence and intensity of heatwaves. The objectives of work are as follow:

- To perform acquisition and pre-processing of data affecting heatwaves at regional and national level.
- To study and implement existing machine learning approaches for heatwave monitoring and prediction.
- To propose and implement an optimal machine learning based solution for heatwave monitoring and prediction, addressing sustainable healthcare.
- To incorporate incremental learning of heatwave patterns and deliver a complete solution for location-wise heatwave monitoring and prediction PAN INDIA for seven upcoming days.
- To evaluate performance of proposed approaches against bench mark algorithms.

1.4 Scope

The scope of this study focuses on predicting heatwaves based on temperature data. The research aims to develop a reliable and accurate heatwave prediction model that utilizes historical temperature records. The scope includes analyzing temperature patterns and variations to identify key indicators of heatwaves. By leveraging temperature data and appropriate Machine Learning techniques, the study seeks to enhance the ability to forecast heatwave occurrences, durations, and intensities for 7 days' lead.

The next section offers an overview of relates work in this area. By examining existing studies, we aim to understand the current landscape and identify areas for further exploration. This related work provides insights into the evolution of research and informs the context for our study.

2. Related Work

The related literature and review of the same can broadly be divided into two parts. Section 2.1 discusses prediction methods which can be used for heatwave prediction. Section 2.2 discusses the existing applications and methods used for heatwave prediction.

2.1 AI Bases Forecasting Methods

Heatwave prediction methods typically rely on a combination of meteorological data, climate models, and advanced statistical or machine learning techniques. Here are some commonly used prediction methods for heatwave forecasting:

2.1.1 Long-Short-Term Memory (LSTM): In 1997, Hochreiter and Schmidhuber introduced a groundbreaking model, known as Long Short-Term Memory (LSTM), which revolutionized many fields. LSTM is a type of recurrent neural network (RNN) architecture extensively utilized in Deep Learning. Its strength lies in its ability to capture long-term dependencies, making it particularly suitable for tasks involving sequence prediction. Unlike traditional neural networks, LSTM incorporates feedback connections, enabling it to process entire sequences of data instead of just individual data points. This feature makes LSTM highly effective in recognizing and forecasting patterns in sequential data such as time series, text, and speech. LSTM has emerged as a powerful tool in artificial intelligence and deep learning, facilitating breakthroughs across various domains by extracting valuable insights from sequential data. The architecture of LSTM addresses the vanishing gradient problem encountered by RNNs. At its core, LSTM functions similarly to an RNN cell. The internal workings of the LSTM network comprise three main components, each responsible for a distinct function, as illustrated in Figure 1 below.

The Logic Behind LSTM: In an LSTM unit, the process unfolds across three distinct parts, often referred to as gates, each playing a crucial role in managing the flow of information: The first part, known as the Forget gate, determines whether the information from the previous timestamp is relevant and should be retained in the memory cell, or if it can be discarded as irrelevant. The second part, termed the Input gate, is responsible for assimilating new information from the input to the cell, enabling the LSTM to learn and adapt based on the current input. The third part, referred to as the Output gate, facilitates the transmission of updated information from the current timestamp to the next, ensuring that relevant information is passed along the LSTM sequence. This entire process, encompassing the three gates, constitutes a single cycle or time step within the LSTM framework. Each LSTM unit can be likened to a layer of neurons in a traditional feedforward neural network, with each neuron possessing a hidden layer and a current state, thereby enabling the model to retain and update information over time.

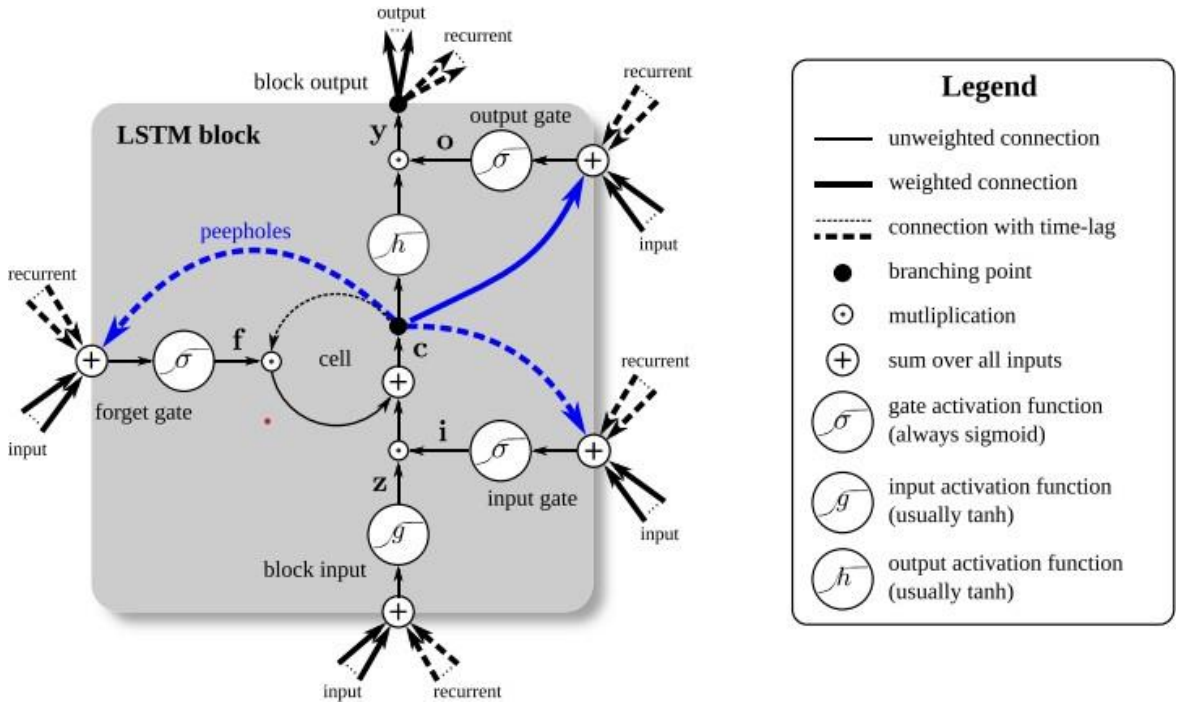

Figure 1: LSTM Architecture [12]

Gates contains sigmoid activations. A sigmoid activation is similar to the tanh activation. Instead of squishing values between -1 and 1, it squishes values between 0 and 1. That is helpful to update or forget data because any number getting multiplied by 0 is 0, causing values to disappears or be "forgotten." Any number multiplied by 1 is the same value therefore that value stay's the same or is "kept." The network can learn which data is not important therefore can be forgotten or which data is important to keep.

Forget gate: The first gate is the forget gate. This gate determines which information should be discarded or retained. Information from the previous hidden state and current input passes through the sigmoid function. Values are outputted between 0 and 1. Closer to 0 means forgetting, and closer to 1 means keeping.

Input Gate: The input gate updates the cell state. First, the previous hidden state and current input are passed into a sigmoid function. This decides which values will be updated, transforming them to lie between 0 and 1. 0 means not important, and 1 means important. The hidden state and current input are also passed into the tanh function, squishing values between -1 and 1 to regulate the network. Then, the tanh output is multiplied with the sigmoid output, where the sigmoid output determines which information from the tanh output is essential to keep.

Cell State: With the gathered information, we can calculate the cell state. The cell state is pointwise multiplied by the forget vector, potentially dropping values in the cell state if multiplied by values close to 0. The output from the input gate is then pointwise added,

updating the cell state with relevant new values found by the neural network. This gives us the new cell state.

Output Gate: Lastly, we have the output gate. This gate determines the next hidden state, which contains information from previous inputs and is also used for predictions. First, the previous hidden state and current input are passed into a sigmoid function. Next, the modified cell state is passed through the tanh function. The tanh output is multiplied by the sigmoid output, deciding the information to be carried by the hidden state. The output is the hidden state. The new cell state and the new hidden state are then carried over to the next time step [12].

2.1.2 Random Forest: The random forest algorithm, proposed by L. Breiman in 2001, has been extremely successful as a general-purpose classification and regression method [13]. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It leverages an ensemble of multiple decision trees to generate predictions or classifications as shown in figure 3. By combining the outputs of these trees, the random forest algorithm delivers a consolidated and more accurate result.

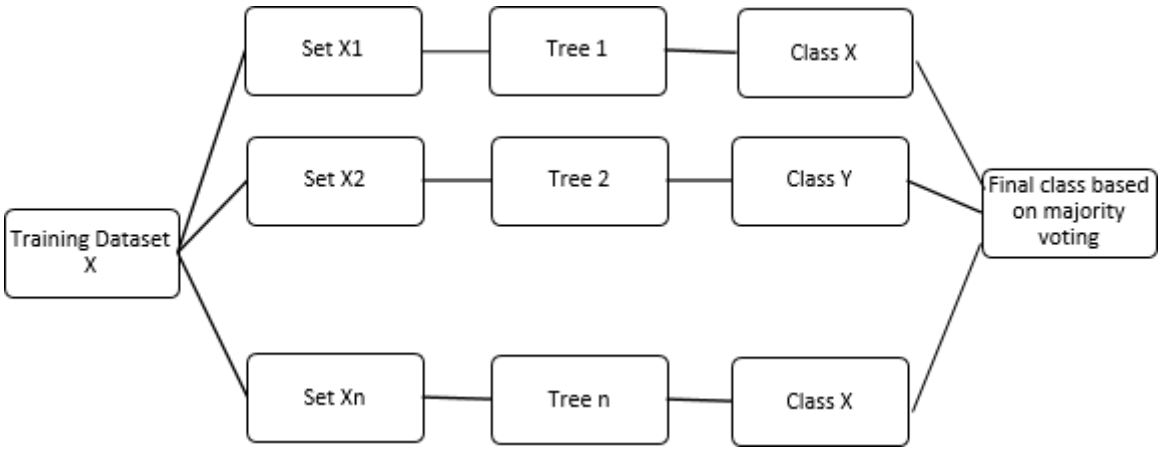

Figure 2: Random Forest Schematic

Its widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems effectively. The algorithm’s strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task.

The Random Forest algorithm's functioning is best understood in the context of ensemble learning in machine learning, where multiple models are combined to make predictions instead

of relying on a single model. Ensemble learning employs two primary methods:

Boosting: This method combines weak learners into strong learners by sequentially creating models in a way that maximizes the final model's accuracy. The subsequent models focus on correcting the errors made by the previous ones. Examples of boosting algorithms include ADA BOOST and XG BOOST.

Bagging: In this method, multiple subsets are created from the training data by sampling with replacement. Each subset is used to build a different model, and the final prediction is determined through majority voting among the individual models. An example of this approach is the Random Forest algorithm. Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forest. Bagging chooses a random sample/random subset from the entire data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently, which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting, is known as aggregation [14].

2.1.3 Support Vector Machine (SVM): SVM is a robust supervised algorithm known for its effectiveness on smaller datasets, especially those that are complex. SVM, short for Support Vector Machine, can handle both regression and classification tasks, though it excels particularly in classification problems. It gained significant popularity during the 1990s when it was first introduced and has since remained a preferred choice for high-performance algorithms with minimal tuning.

In the SVM algorithm, each observation is plotted as a point in an n-dimensional space, where n represents the number of features in the dataset. The main objective is to find an optimal hyperplane that effectively classifies the data points into their respective classes. Before delving into SVM's working, let's first comprehend the two fundamental terms used in the algorithm: "The support vector" and "Hyper-Plane."

Hyper-Plane: The hyperplane acts as a decision boundary that separates the two classes in SVM. Data points falling on either side of the hyperplane belong to different classes. The dimension of the hyperplane depends on the number of input features in the dataset. For instance, if there are two input features, the hyper-plane will be a line, while for three features, it becomes a two-dimensional plane.

Support-Vectors: Support vectors are the data points closest to the hyperplane, and they influence the hyperplane's position and orientation. The goal is to choose a hyperplane for which the margin, i.e., the distance between support vectors and the hyperplane, is maximized. Even a slight alteration in the position of these support vectors can lead to a change in the hyperplane.

The best way to understand the SVM algorithm is by focusing on its primary type, the SVM classifier. The idea behind the SVM classifier is to come up with a hyper-lane in an N-dimensional space that divides the data points belonging to different classes. However, this hyper-pane is chosen based on margin as the hyperplane providing the maximum margin between the two classes is considered. These margins are calculated using data points known as Support Vectors. Support Vectors are those data points that are near to the hyper-plane and help in orienting it.

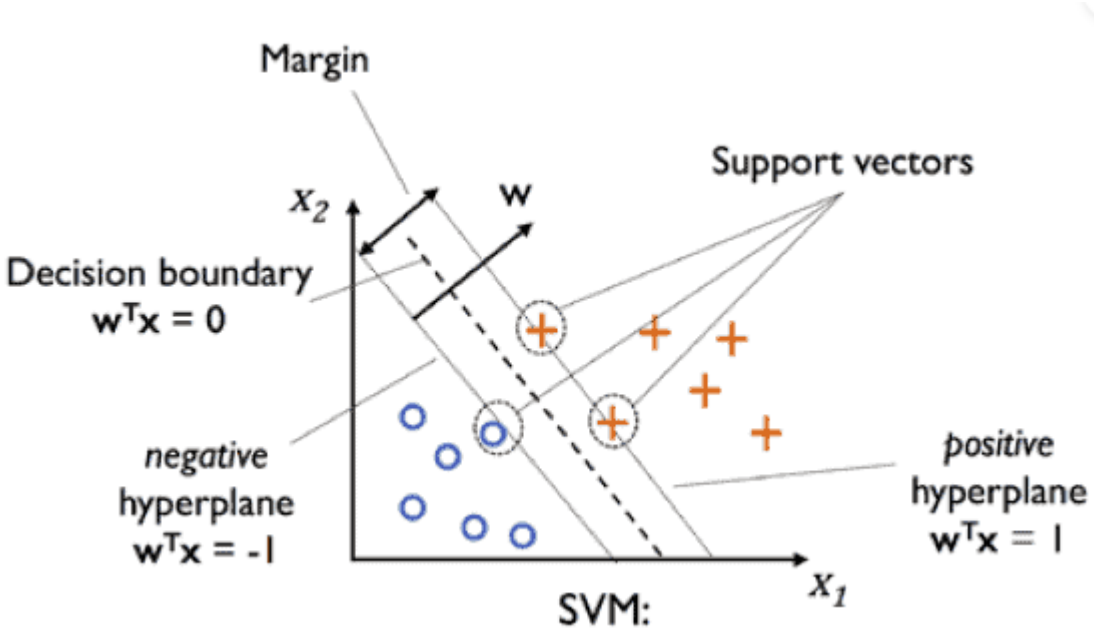

Figure 3: Main Concept of SVM [15]

There exist two types of Support Vector Machines:

Linear SVM or Simple SVM: Linear SVM is employed when dealing with linearly separable data. If a dataset can be categorized into two classes using a single straight line, it is considered linearly separable data, and the classifier used is called the linear SVM classifier. This type of SVM is typically utilized for linear regression and classification tasks.

Nonlinear SVM or Kernel SVM: Nonlinear SVM comes into play when handling nonlinearly separated data, which means a dataset that cannot be classified using a straight line. In this scenario, a classifier known as a nonlinear SVM classifier is used. It provides greater flexibility for nonlinear data by incorporating additional features to accommodate fitting a hyperplane within a multi-dimensional space, moving beyond the constraints of a two-dimensional space [15].

2.2 Literature Survey

In the context of heatwave prediction, several related works have been conducted to explore various methodologies and techniques as given below.

2.2.1 International Status:

Osman et al. [16] evaluates the prediction skill of two sub seasonal forecast models, POAMA and BCC-CPS, regarding temperature and circulation anomalies during the December 2013 heatwave. It compares the performance of both models, highlighting POAMA's superior skill for longer lead times and BCC-CPS's better performance for shorter leads. The analysis includes the probabilistic approach of the models in predicting temperature anomalies, with POAMA consistently showing a higher chance of exceeding the upper tercile. The study suggests employing a multi-model ensemble strategy to enhance surface temperature forecasting during heatwaves. Methods involved evaluating prediction skill, considering reforecasts, computing weekly ensemble-mean forecasts, and comparing model performance in terms of warm anomalies. Results indicate moderate skill for POAMA at longer lead times and better performance of BCC-CPS for shorter leads, particularly in locating anomaly maxima during week 3 of the heatwave. Additionally, POAMA's probabilistic approach predicts a higher chance of temperature anomalies exceeding the upper tercile, and circulation anomaly forecasts aid in understanding anomaly locations.

Russo et al. [17] compare the intensity of heatwaves across continents and historical periods. It utilizes daily maximum temperature data to identify and classify the most severe heatwaves globally over three distinct study periods. The HWMI is then applied to project the future frequency and severity of heatwaves under different climate scenarios. Using ensemble outputs from the Coupled Model Intercomparison Project Phase 5 (CMIP5), the study forecasts the occurrence and severity of heatwaves under various Representative Concentration Pathways (RCPs). To ensure consistency, all CMIP5 models and reanalysis data are interpolated onto the same T63 grid. Heatwave severity is classified based on a new scale ranging from Normal to Ultra Extreme, as outlined in Appendix B of the study. The results reveal a concerning trend: the percentage of global area affected by moderate, severe, and extreme heatwaves have tripled between 2002 and 2012 compared to previous periods. Notably, the 2010 Russian heatwave emerges as the most intense globally in recent decades, surpassing even the 2003 European summer heatwave in amplitude and spatial extent. Model predictions suggest a significant increase in the likelihood of extreme and very extreme heatwaves in the future, especially by the end of the century. Under the most severe IPCC AR5 scenario, regions like southern Europe could experience events comparable to the 2010 Russian heatwave every two years. Despite some limitations in accurately simulating individual extreme heatwave events, general circulation models (GCMs) exhibit consistency in capturing global and regional heatwave trends, as validated against reanalysis data. The study also provides a comprehensive list of severe heatwaves occurring in various regions, including the U.S., Europe, and Russia, each characterized by different magnitudes and spatial extents.

Ford et al. [18] introduces a method for evaluating heatwave forecasts across different timescales, specifically focusing on the NCEP coupled forecast model. Through this evaluation, the study aims to identify sources of poor forecast skill within the model. Key methods employed in the paper include the development of an extreme heat validation approach tailored for medium-range forecast models, such as the NCEP model. This approach utilizes a weighting strategy based on the Poisson function to seamlessly transition from short-term day-by-day

forecasts to expanding time means across sub seasonal timescales. The evaluation of heatwave forecast skill involves the use of various deterministic and probabilistic skill metrics. Additionally, the study delves into the analysis of land-atmosphere feedbacks within the NCEP model to understand their impact on heatwave forecasts, particularly focusing on regions like the Southwest United States. Results from the study indicate that while the choice of skill metric does not significantly affect the overall skill of heatwave forecasts, there are distinct spatial patterns in forecast skill based on different temperature metrics. The NCEP model exhibits deficiencies in accurately persisting observed heatwaves, particularly in the Southwest United States, and misrepresents land-atmosphere feedbacks. The Poisson weighting strategy, coupled with a range of skill metrics, proves to be a versatile framework for validating dynamical model heatwave forecasts at sub seasonal timescales. Evaluation using the Equitable Threat Score (ETS) metric highlights a decrease in forecast reliability with longer lead times, with under-forecasting of heatwaves in the Midwest contributing to this decline.

Asadollah et al. [19] develop physical-empirical models using machine learning algorithms to predict the number of annual heatwave days in different regions of Iran using synoptic predictors. The study aims to identify the optimal parameter combination as the predictors of heatwaves by examining the effects of numerous weather components. The proposed hybrid approach, ABR-DT, showed better performance in predicting heatwave days compared to the other two alternatives. The paper uses two classical machine learning algorithms, namely decision tree (DT) and random forests, and a novel hybrid technique consists of Ada-Boost Regression and decision tree (ABR-DT) for forecasting annual heatwave days (HWDs) of Iran from synoptic predictors. The Principal Component Analysis was employed to reduce the diverse predictors and their features to the most optimal structure. The paper found that the proposed hybrid forecasting approach, ABR-DT, is effective in predicting heatwave days in different regions of Iran using synoptic predictors. The ABR-DT approach showed better performance in predicting heatwave days compared to the other two alternatives, with a correlation coefficient (CC) of 0.860 and mean absolute error (MAE) of 6.929, using only specific humidity and wind component as predictors. The spatial performance indices over eight different climate regions of Iran also showed the better performance of ABR-DT, which improved the CC and MAE of its two alternatives up to 185 and 19%. Open issues include that paper focuses solely on predicting heatwaves in Iran and may not be generalizable to other regions or climates and does not provide a detailed analysis of the physical mechanisms underlying heatwaves in Iran, which could limit the interpretability and applicability of the proposed machine learning models.

French et al. [20] propose a statistical framework that combines extreme value theory and functional data analysis to compute probabilities of rare heatwaves and other extreme weather events. The approach can be applied to data from the North American Regional Climate Change Assessment Program to compute probabilities for heatwaves of any pre-specified temporal duration, spatial extent, and overall magnitude. The paper aims to provide a flexible and readily applicable framework that allows researchers to compute probabilities of extreme weather events. The paper proposes a statistical framework that combines extreme value theory (EVT) and functional data analysis (FDA) to compute probabilities of rare heatwaves and other extreme weather events. The approach allows for the computation of probabilities for heatwaves of any

pre-specified temporal duration, spatial extent, and overall magnitude. Open issues include that the proposed framework is based on computer model predictions, which may have some limitations and uncertainties and it assumes that the data are stationary, which may not be true in practice. The framework is focused on North America and does not take into account the potential impact of climate change on extreme weather events, which may affect the accuracy of the predictions.

Chantry et al. [21] discuss the use of machine learning techniques to improve weather and climate prediction. it highlights the potential benefits of using machine learning techniques to improve weather and climate prediction. This could lead to more accurate and reliable forecasts, which could have a significant impact on a range of industries, including agriculture, transportation, and energy. Additionally, the paper identifies some of the challenges and limitations of using machine learning in this context, which could help guide future research and development in the field. There are different types of machine learning techniques that can be used in this context, including hard, medium, and soft AI. Hard AI involves using algorithms to make predictions based on historical data, while medium AI involves using machine learning to identify patterns and relationships in the data. Soft AI involves using machine learning to simulate human decision-making processes. There are several challenges and limitations to using machine learning in weather and climate modelling, including the need for large amounts of high-quality data, the potential for overfitting, and the difficulty of interpreting the results. Despite these challenges, there is significant potential for machine learning to improve weather and climate prediction, and further research in this area is needed to fully realize these benefits.

Domeisen et al. [22] discuss the threat of heatwaves to human health and ecosystems, and the need for prediction to minimize adverse impacts. They document current capabilities for heatwave prediction and outlines projected changes under anthropogenic warming. The paper also outlines the drivers of, and feedbacks associated with, heatwaves, and provides an overview of heatwave prediction on timescales of several days to decades, followed by heatwave projection in future climates. The focus is on land heatwaves, with an overview of marine heatwaves and their impacts available elsewhere. The paper highlights several open issues that require further research to improve heatwave prediction and projection. These include: (i) Improving the representation of atmospheric dynamics, atmospheric and soil moisture, and surface cover in models to better understand the relevant drivers and feedbacks associated with heatwaves. (ii) Developing better methods for predicting heatwaves beyond timescales of a few weeks, as current predictions are challenging beyond this timeframe. (iii) Improving our understanding of the drivers of heatwaves in different regions, as the driving mechanisms and relative importance of these mechanisms vary depending on the location. (iv) Developing better methods for predicting the occurrence of humid heatwaves, which are becoming more frequent in some regions, especially southern Asia.

Jacques-Dumas et al. [23] explore the use of Deep Learning architectures as alternative strategies to predict extreme heatwaves occurrences from a very limited amount of available relevant climate data. The study aims to address issues such as the aggregation of climate data of different natures, the class-size imbalance that is intrinsically associated with rare event prediction, and the potential benefits of transfer learning to address the nested nature of extreme events. The

study uses Convolutional Neural Network-based Deep Learning frameworks, with large-class under sampling and transfer learning to achieve significant performance in forecasting the occurrence of extreme heatwaves, at three different levels of intensity. The study used a limited amount of available relevant climate data, and the findings may not be generalizable to other datasets.

McElroy et al. [24] discuss the classification of heat events and their association with hospitalizations in the County of San Diego. The researchers aimed to identify the criteria for activating early warning systems and to quantify the discrepancies in heat event classifications based on local meteorological data. The goal of this paper was to identify optimal definitions of heatwaves and extreme heat events using sub-regional meteorological data to maximize the benefits of early warning systems to population health. The researchers used local meteorological data to create 18 event classifications using various combinations of temperature metric, percentile, and duration for both county-level and climate zone level meteorological data within San Diego County. They then utilized propensity score matching and Poisson regressions to ascertain the association between heatwave exposure and risk of hospitalization for heat-related illness and dehydration for the 1999–2013 period. The researchers estimated both relative and absolute risks for each heat event classification in order to identify optimal definitions of heatwaves and extreme heat events for the whole city and in each climate zone to target health impacts. The researchers found that heat-related illness differs vastly by level (county or zone-specific), definition, and risk measure. They also found that the county-level definitions were systematically biased when compared to climate zone definitions with the largest discrepancy of 56 attributable hospitalizations. The researchers concluded that definitions based on county-level defined thresholds do not provide an accurate picture of the observed health effects and will fail to maximize the potential effectiveness of heat warning systems. Absolute rather than relative risks are a more appropriate measure to define the set of criteria to activate early warnings systems and thus maximize public health benefits.

The study provides insights into the discrepancies between county-level and climate zone definitions of heatwaves and extreme heat events. The study was limited to the County of San Diego and may not be generalizable to other regions. The study only focused on hospitalizations for heat-related illness and dehydration and did not consider other health outcomes. The study did not consider the socioeconomic and demographic factors that may influence the health impacts of heat waves and extreme heat events. Further research is needed to identify optimal definitions of heat waves and extreme heat events in other regions. Future studies should consider other health outcomes and the socioeconomic and demographic factors that may influence the health impacts of heat waves.

Kim et al. [25] introduce a method for predicting temperatures using LSTM, aiming to offer short-term forecasting insights for heatwaves in Korea. Such forecasts could potentially aid in preventing or lessening the adverse effects of heatwaves on human health, infrastructure, and the environment. The study proposes utilizing LSTM (Long Short-Term Memory), a type of recurrent neural network (RNN) renowned for its ability to capture long-term dependencies. LSTM has found widespread application across diverse domains, including natural language processing, speech recognition, and time series prediction. The effectiveness of the proposed methodology was assessed by comparing the predicted temperatures with actual field

measurements. Accuracy was gauged by computing the RMSE (Root Mean Square Error) across various temperature ranges. Despite the promising results, several unresolved issues persist. These include the necessity to validate the proposed methodology across broader spatial regions or different provinces, utilizing appropriate datasets to assess its applicability and robustness. Additionally, comparative analyses with other machine learning models and numerical forecasting methods are warranted to elucidate the strengths and limitations of different approaches.

Ding et al. [26] study the statistical characteristics of heatwave precursors in China and model prediction. The paper aims to identify the precursors of heatwaves in China and predict them using the 250hPa regional-scale geopotential height anomalies. The paper uses statistical analysis and modeling approaches to study the precursors of heatwaves in China. The statistical analysis is based on the 250hPa regional-scale geopotential height anomalies, which are used to identify the precursors of heatwaves. The modeling approach is based on physical decomposed ECMWF forecast data at 500hPa, which is used to predict the occurrence of heatwaves. The study finds that among 87 events of regional dry heatwave (HW) in China during 1979—2008, precursors can be found from 83 regional dry HW events by the 250hPa regional-scale geopotential height anomalies. The height anomalies move westward in the low latitudes and eastward in the middle latitudes, with an average lead of 7 days before HW events occurred. 19% of the precursors for regional dry HW in southern China originate in the northwest Pacific, 4.6 days in advance of the HW on average, while most of the precursors (81%) start from Europe to Northwest China with a lead of 2-15 days, then propagate in the middle latitudes to 100-110°E and 40°N. Open issues includes that the study focuses on regional dry heatwaves in China and does not consider other types of heatwaves or other regions. Additionally, the study is based on statistical analysis and modeling, which may have limitations and uncertainties.

2.2.2 National Status

Dube et al. [27] analyze the ability of the operational global ensemble prediction system (EPS) at the National Centre for Medium Range Weather Forecasting (NCMRWF, NEPS) in predicting the probabilities of the maximum 2 m temperature (TMAX) in India during March to May 2019. The paper aims to demonstrate improved skill in probabilistic forecasts of TMAX over India during MAM 2019. The paper uses two approaches to correct the bias in the mean and spread of the ensemble prediction system (EPS) forecasts of maximum 2 m temperature (TMAX) in India during March to May 2019. The first approach is the Decaying Average technique, which corrects the bias in the mean (1st moment) of the TMAX forecasts. The second approach is the Variation Inflation method, which corrects the bias in the spread (2nd moment) of the TMAX forecasts. The paper then compares the forecasts of TMAX from the raw and the bias-corrected data using standard verification metrics for probabilistic forecasts like Brier Score (BS), Brier Skill Score (BSS), ROC and Reliability Diagrams as well as the Value score. The findings of this paper suggest that the bias correction in the mean (1st moment) of the TMAX forecasts using the Decaying Average technique shows a large improvement in terms of reduced false alarms and increased hit rate. The study focuses on the probabilistic forecasts of maximum 2 m temperature (TMAX) over India during March to May 2019. Therefore, the findings of this study may not be generalizable to other regions or time periods. Additionally, the paper does not

address the potential impact of climate change on the accuracy of the probabilistic forecasts of TMAX.

Naveen et al. [28] analyze the spatial distribution, frequency, and trend of heatwaves over India using gridded daily temperature data sets for the period 1951-2019. The study aims to identify the number of heatwave days and their spatial distribution in India during the summer months of April and May. The paper uses gridded daily temperature data sets for the period 1951-2019 to compute the arithmetic mean (AM), standard deviation (SD), coefficient of variation (CV), and trends of monthly maximum temperature. The number of heatwave days were identified using the criteria given by India Meteorological Department (IMD). The analysis was confined to the two summer months of April and May only. The study uses statistical and spatial analysis approaches to identify the characteristics of heatwaves over India. The paper finds that the core hot region with temperatures exceeding 40 °C lies over central India extending towards the northwest. The spatial distribution of the arithmetic mean (AM) shows higher values during May. The standard deviation (SD) distribution shows higher values over the northeast of central India decreasing towards the southwest. The coefficient of variation (CV) distribution shows higher values over the north decreasing toward the south. Limitation of the study is that it only analyzes the two summer months of April and May, and the findings may not be applicable to other months of the year.

Chand et al. [29] design and evaluate Artificial Neural Networks (ANN) to predict daily Steadman Heat Index (SHI) using temperature and humidity in Australia. The paper aims to develop a suitable intelligent data analytic tool for weather prediction, climate change studies, and probable evaluation of dry climatic conditions in the near future. The performance of the model is assessed using various metrics such as the spread and distribution of predicted SHI, Legates and McCabe Index, Mean Absolute Error, Root Mean Square Error, Coefficient of Determination, the Willmott's Index of Agreement, and Nash–Sutcliffe Coefficient of Efficiency.

Sahai et al. [30] propose a probabilistic early health warning system (EHWS) based on meteorological parameters for forecasting the incidence of malaria and diarrhea in India. The goal of the paper is to develop a method for probabilistic forecasting of disease incidences in extended range time scale (2-3 weeks in advance) over India based on an unsupervised pattern recognition technique that uses meteorological parameters as inputs. The paper uses the Self-Organizing Map (SOM) technique to cluster the meteorological parameters and disease incidences. The findings of the paper suggest that the increased probabilities of high (less) rainfall, high (low) minimum temperature and low (moderate) maximum temperature are more (less) conducive for both diseases over the studied locations, but have different thresholds. The EHWS is found to be a useful tool with reasonable skill to provide the climate-health outlook about possible disease incidence at least 2 weeks in advance for any location or grid over India. The advantages of the paper include the development of a new EHWS for the real-time prediction of disease incidences over the Indian region by using the weather parameters as predictors. The open issues include the need for linking health and environmental information across high spatial and temporal resolutions using historical and real-time surveillance information to inform the development of these early warning systems to save lives.

3. Proposed Methodology

The proposed approach for heatwave prediction involves the following steps as highlighted in figure 4.

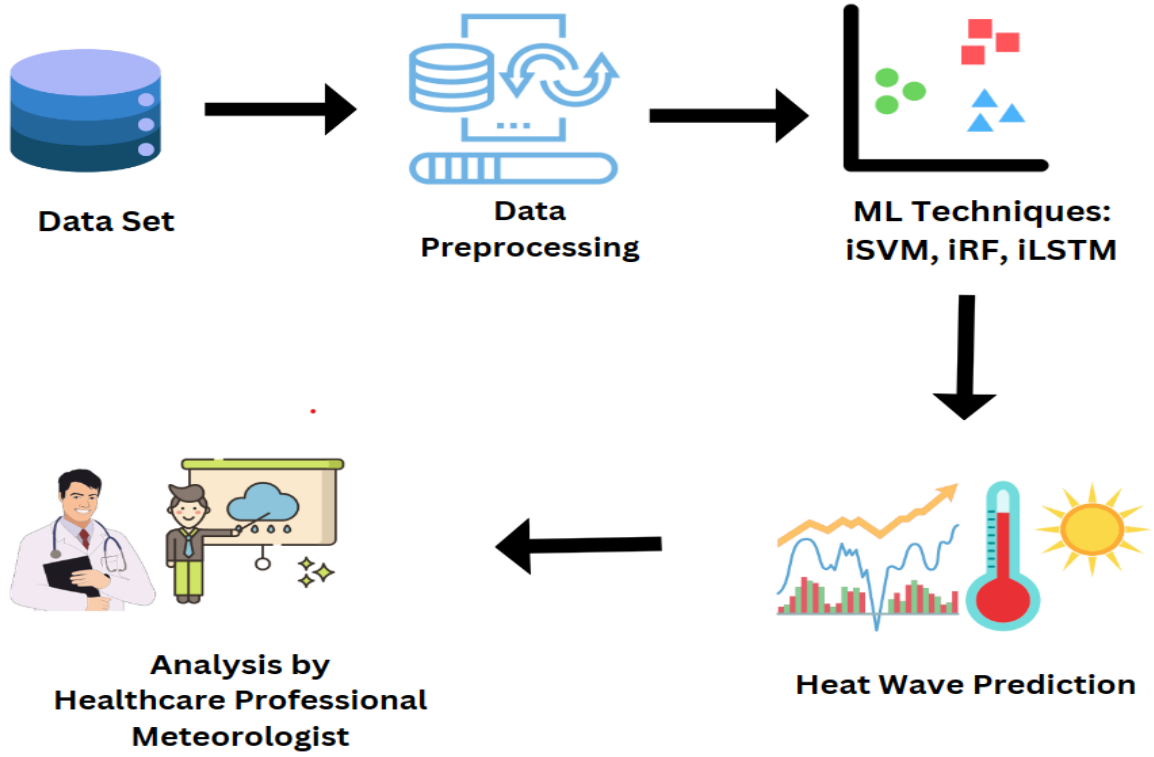

Figure 4: Proposed Approach for heat wave prediction

Data Collection & Pre-processing: The research data is sourced from the India Meteorological Department, Pune, retrieved in real-time from their website at https://imdpune.gov.in/lrfindex.php. This data forms the basis for constructing the heatwave prediction model. Subsequently, the data undergoes cleaning and preprocessing procedures to address issues such as missing values, outliers, and noise in the temperature data.

Incremental Machine Learning Model Development: During the model selection phase, we explore various prediction models for heatwave forecasting, incorporating both traditional Machine Learning algorithms and deep learning techniques. This includes Support Vector Machines (SVM), Random Forest, Long Short-Term Memory (LSTM) networks, and techniques for incremental learning. SVM is considered for its effectiveness in handling linearly separable data, while Random Forest is chosen for its ability to handle complex datasets. LSTM networks are particularly suitable for capturing spatio-temporal patterns in temperature data. Additionally, techniques for incremental learning are explored, allowing the model to adapt and learn from new data continuously. Through comprehensive performance evaluation, we aim to identify the most suitable model for accurate and effective heatwave prediction based on historical temperature records.

Prediction: During the model training phase, historical temperature data with labeled heatwave events is utilized to train each model. Once trained, these models are deployed to predict

heatwave occurrences in the future based on new temperature data. Heat watches are issued when there's a likelihood of excessive heat in the next 24 to 72 hours. These watches serve as a precautionary measure when the risk of a heatwave is elevated. A hea wave is defined as a period of unusually hot weather, typically lasting two or more days. According to the criteria provided by the Indian Meteorological Department, a heatwave is considered if the maximum temperature of a station reaches at least 40°C or more for Plains and at least 30°C or more for Hilly regions.

Analysis: Heatwave prediction offers numerous advantages across various domains. Healthcare professionals benefit from enhanced preparedness, increased public awareness, and the ability to identify vulnerable populations at risk during heatwaves. Meteorologists gain the capability to issue timely alerts and provide improved weather forecasting. Governments can plan and implement effective emergency care measures and public health interventions to mitigate the impacts of heatwaves. Overall, heatwave prediction contributes to safeguarding public health and enhancing societal resilience to extreme weather events.

4. Expected Contribution & Future Plan

This research aims to deepen the understanding of the meteorological drivers of heatwaves and develop a robust heatwave prediction model based on meteorological data. By incorporating AI technologies, such as Support Vector Machines, Random Forest, and Long Short-Term Memory networks, the research seeks to develop a robust heatwave prediction model based on comprehensive meteorological data. Strategies for integrating AI into heatwave prediction systems will be explored, and advanced AI techniques will be optimized to address prediction-related challenges.

Expected Contribution of this work:

- The research aims to design incremental machine learning approaches to predict heatwave intensity and location accurately for next seven days.
- The research aims to pre-empt health emergencies by performing heatwave prediction in advance.
- The research aims to enable decision making by meteorologist for timely prediction.

This research aims to contribute to more accurate and reliable heatwave forecasts, enabling timely and effective response measures to mitigate the adverse impacts of heatwaves. By leveraging AI and meteorological data, aim of this research is to better understand and predict heatwaves.
The research begins by weather data acquisition and preprocessing. Study, exploration and implementation of existing techniques will be done to build a strong foundation. Based on this, a detailed research approach is developed, serving as a roadmap for experiments and analysis. Incremental learning will be incorporated and performance of proposed approach will be evaluated against benchmark algorithms. Finally, the Ph.D. thesis will be written, documenting the research findings, and conclusions. The future plan is to implement proposed approach for heatwave prediction as per given timeline in figure 5.

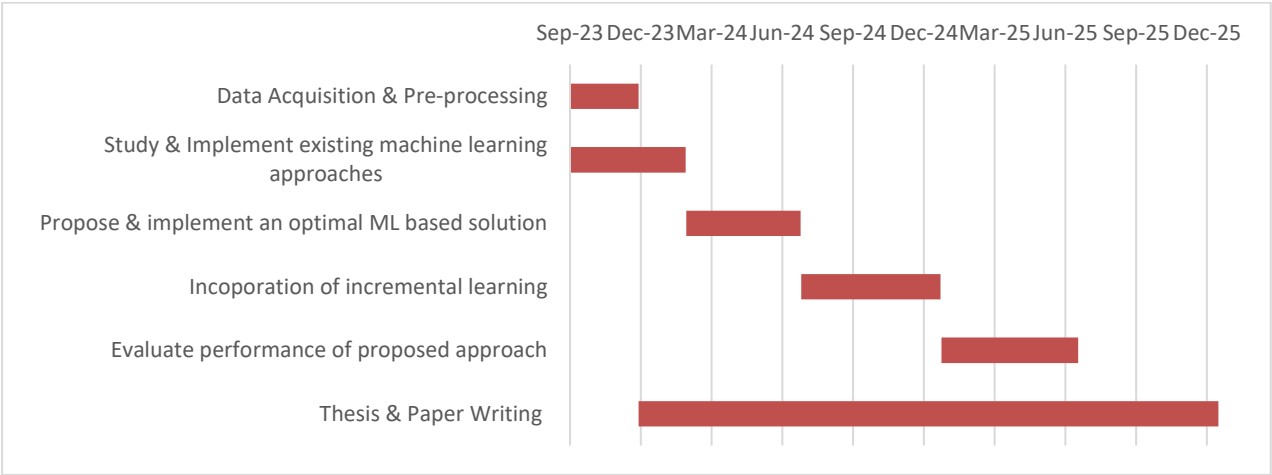

Figure 5: Timeline Chart

References

[1] YT Wen, TT Liu, T Lin, YF Lin, CC Chen, WM Kung, CC Huang, TJ Lin, YH Wang, L Wei (2015). Heatstroke effect on brain heme oxygenase-1 in rats. International journal of medical sciences, 12(9), pp 737-41. https://doi.org/10.7150/ijms.12517

[2] Global Warming of 1.5 ºC — (ipcc.ch) accessed on 28th July 2023

[3] https://tcd.uchicago.edu/insight/report-on-climate-change-and-heat-induced-mortality-in-India accessed on 28th July 2023.

[4] https://economictimes.indiatimes.com/news/politics-and-nation/1-5-million-more-people-may-die-in-india-by-2100-due-to-extreme-heat-by-climate-change-study/articleshow/71842685.cms?mc_cid=dec4abb97e&mc_eid=89ee108d17 accessed on 28th July 2023

[5] https://ghhin.org/understanding-heat/ accessed on 31st July 2023

[6] J Zuo, S Pullen, J Palmer, H Bennetts, N Chileshe, T Ma, Impacts of heat waves and corresponding measures: a review, Journal of Cleaner Production, Volume 92,2015, pp 1-12, https://doi.org/10.1016/j.jclepro.2014.12.078.

[7] M Andrea, Y Zhang, C Carrozzo, E Bompard, G Chicco, E Roggero, and G Galofaro., "Evaluation of the Impact of Heat-Wave on Distribution System Resilience," 2021 International Conference on Smart Energy Systems and Technologies(SEST), Vaasa, Finland, 2021, pp. 1-6, https://doi.org/10.1109/SEST50973.2021.9543201.

[8] AK Sahai, R Mandal, S Joseph, S Saha, P Awate, S Dutta, A Dey, RChattopadhyay, R Phani, DR Pattanaik, S Despande, (2020). Development of a probabilistic early health warning system based on meteorological parameters. Scientific reports, 10(1), 14741.

[9] S Secinaro, D Calandra, A Secinaro, V Muthurangu, P Biancone, (2021). The role of artificial intelligence in healthcare: a structured literature review. BMC medical informaticsand decision making, 21, pp 1-23.

[10] B Chand, T Nguyen-Huy, & R C Deo (2021) Artificial neural networks for predictionof Steadman Heat Index Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation: Theory and Practice of Hazard Mitigation, pp 293-357

[11] S Hochreiter & J Schmidhuber, (1997) Long short-term memory Neural computation, 9(8), pp 1735-1780

[12] K Greff, R K Srivastava, J Koutník, B R Steunebrink, & J Schmidhuber, (2016) LSTM: A search space odyssey IEEE transaction on neural networks and learning systems, 28(10), 2222-2232

[13] L Breiman (2001) Random forests Machine learning, 45, pp 5-32

[14] K Fawagreh, M M Gaber, & E Elyan, (2014) Random forests: from early developments to recent advancements Systems Science & Control Engineering: An Open Access Journal, 2(1), pp 602-609 https://doiorg/101080/216425832014956265

[15] M A Chandra, & S S Bedi, (2021) Survey on SVM and their application in image classification International Journal of Information Technology, 13, pp 1-11

[16] M Osman, MS Alvarez, Subseasonal prediction of the heat wave of December 2013 in Southern South America by the POAMA and BCC-CPS models Clim Dyn 50, 67–81 (2018) https://doiorg/101007/s00382-017-3582-4

[17] S Russo, A Dosio, R G Graversen, J Sillmann, H Carrao, M B Dunbar, & J V Vogt (2014) Magnitude of extreme heat waves in present climate and their projection in a warming world Journal of Geophysical Research: Atmospheres, 119(22), 12-500

[18] T W Ford, P A Dirmeyer, & D O Benson (2018) Evaluation of heat wave forecasts seamlessly across subseasonal timescales NPJ Climate and Atmospheric Science, 1(1), 20

[19] S B H S Asadollah, N Khan, A Sharafati, S Shahid, E S Chung, & X J Wang (2021) Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models Stochastic environmental research and risk assessment, Elsevier, pp 1-16 https://doiorg/101007/s00477-021-02103-z

[20] J French, P Kokoszka, S Stoev, & L Hall (2019) Quantifying the risk of heat waves using extreme value theory and spatio-temporal functional data Computational Statistics & Data Analysis, Elsevier, 131, pp 176-193 https://doiorg/101016/jcsda201807004

[21] M Chantry, H Christensen, P Dueben, & T Palmer (2021) Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI Philosophical Transactions of the Royal Society A, 379(2194), 20200083 https://doiorg/101098/rsta20200083

[22] D I Domeisen, E A Eltahir, E M Fischer, R Knutti, S E Perkins-Kirkpatrick, C Schär, SI Seneviratne, A Weisheimer, & H Wernli (2023) Prediction and projection of heatwaves Nature Reviews Earth & Environment, 4(1), pp 36-50 https://doiorg/101038/s43017-022-00371-z

[23] V Jacques-Dumas, F Ragone, P Borgnat, P Abry, & F Bouchet (2022) Deep learning-based extreme heatwave forecast Frontiers in Climate, 4

[24] S McElroy, L Schwarz, H Green, I Corcos, K Guirguis, A Gershunov, & T Benmarhnia (2020) Defining heat waves and extreme heat events using sub-regional meteorological data to maximize benefits of early warning systems to population health Science of the Total Environment, Elsevier, 721 https://doiorg/101016/jscitotenv2020137678

[25] Y I Kim, D Kim, & S O Lee (2020) Prediction of temperature and heat wave occurrence for summer season using machine learning Journal of Korean Society of Disaster and Security, 13(2), pp27-38 https://doiorg/1021729/ksds202013227

[26] T Ding, & W H Qian (2012) Statistical characteristics of heat wave precursors in China and model prediction Chinese Journal of Geophysics, 55(5), pp 1472-1486 https://doiorg/106038/jissn0001-5733201205005

[27] A Dube, H Singh, & R Ashrit (2021) Heat waves in India during MAM 2019: Verification of ensemble based probabilistic forecasts and impact of bias correction Atmospheric research, Elsevier 251, 105421 https://doiorg/101016/jatmosres2020105421

[28] N Naveena, G C Satyanarayana, A D Raju, K S Rao, & N Umakanth (2021) Spatial and statistical characteristics of heat waves impacting India AIMS Environmental Science, 8(2), pp 117–134 https://doiorg/103934/environsci2021009

[29] B Chand, T Nguyen-Huy, & R C Deo (2021) Artificial neural networks for prediction of Steadman Heat Index Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation: Theory and Practice of Hazard Mitigation, Springer, pp 293-357 https://doiorg/101007/978-981-15-5772-9_16

[30] A K Sahai, R Mandal, S Joseph, S Saha, P Awate, S Dutta, A Dey, R Chattopadhyay, R Phani, D R Pattanaik, & S Despande (2020) Development of a probabilistic early health warning system based on meteorological parameters Scientific Reports, 10(1) https://doiorg/101038/s41598-020-71668-6

[31] B B Goswami, V Venugopal, D Sengupta, M S Madhusoodanan, P K Xavier (2006) Increasing trend of extreme rain events over India in a warming environment Science, 314(5804), pp1442-1445

[32] C Brimicombe, C H B Lo, F Pappenberger, C Di Napoli, P Maciel, T Quintino (2023) Wet Bulb Globe Temperature: Indicating extreme heat risk on a global grid GeoHealth, 7, e2022GH000701 https://doiorg/101029/2022GH000701

[33] V Rodriguez-Galiano, M Sanchez-Castillo, M Chica-Olmo, & M J O G R Chica-Rivas (2015) Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines Ore Geology Reviews, Elsevier, 71, pp 804-818

[34] F A Gers, J Schmidhuber, & F Cummins (2000) Learning to forget: Continual prediction with LSTM Neural computation, 12(10), pp 2451-2471